Speech Intelligibility Assessment with Uncertainty-Aware Whisper Embeddings and sLSTM

Ryandhimas E. Zezario* and Dyah A.M.G. Wisnu† and Hsin-Min Wang‡ and Yu Tsao*
* Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: {ryandhimas,yu.tsao}@citi.sinica.edu.tw
† Social Network and Human Centered Computing, Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: dyahayumgw@iis.sinica.edu.tw
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: whm@iis.sinica.edu.tw

***Abstract*—Non-intrusive speech intelligibility prediction remains challenging due to variability in speakers, noise conditions, and subjective perception. We propose an uncertainty-aware approach that leverages Whisper embeddings in combination with statistical features—specifically, the mean, standard deviation, and entropy computed across the embedding dimensions. The entropy, computed via a softmax over the feature dimension, serves as a proxy for uncertainty, complementing global information captured by the mean and standard deviation. To model the sequential structure of speech, we adopt a scalar long short-term memory (sLSTM) network, which efficiently captures long-range dependencies. Building on this foundation, we propose iMTI-Net, an improved multi-target intelligibility prediction network that integrates convolutional neural network (CNN) and sLSTM components within a multitask learning framework. It jointly predicts human intelligibility scores and machine-based word error rates (WER) from Google ASR and Whisper. Experimental results show that iMTI-Net outperforms the original MTI-Net across multiple evaluation metrics, demonstrating the effectiveness of incorporating uncertainty-aware features and the CNN-sLSTM architecture.**

I. Introduction

Speech intelligibility is an essential indicator for evaluating a wide range of speech-related applications, including speech enhancement [1], [2], hearing aid (HA) devices [3], [4], and telecommunications [5]. The direct measurement is based on human listening tests, where listeners are required to recognize words from the played speech samples. The ratio of correctly recognized words to the total number of words is used to determine the intelligibility score. However, despite the reliability of human listening tests, a sufficient number of listeners is necessary to obtain unbiased measurements. Furthermore, this requirement limits the practicality and scalability of human evaluations. To overcome this problem, a series of signal processing-based approaches have been proposed, such as the articulation index (AI) [6], speech intelligibility index (SII) [7], extended SII (ESII) [8], speech transmission index (STI) [5], and short-time objective intelligibility (STOI) [9]. However, most of these methods require ground-truth references to produce reliable evaluation scores.

With the growing interest in reliable non-intrusive speech intelligibility methods, where ground-truth references are not necessarily required for evaluation, deep learning models have received increasing attention. This trend is further supported by the availability of large-scale datasets labeled by human annotators with corresponding speech intelligibility scores, enabling the development of deep learning-based intelligibility assessment models [10]–[20]. In the early stages of this development, most well-known approaches [10]–[13] relied on traditional speech processing techniques to extract acoustic features. More recently, the introduction of large-scale pre-trained speech models, such as self-supervised learning models (e.g., wav2vec [21] and HuBERT [22]) and weakly supervised models (e.g., Whisper [23]), has further advanced the field by providing richer and more generalizable acoustic features [16]–[19], [24]. Despite these advancements, developing accurate and generalizable intelligibility assessment models remains challenging. The subjective nature of intelligibility, variations in speaker characteristics, and environmental conditions can degrade prediction performance. Additionally, capturing long-range dependencies in temporal speech patterns is non-trivial, particularly under noisy or mismatched conditions.

To address these issues, we propose an uncertainty-aware approach that combines Whisper embeddings with statistical features. Specifically, we extract the mean, standard deviation, and entropy of the Whisper embeddings across the feature dimension. We hypothesize that the mean and standard deviation capture global characteristics of the embeddings over time, while the entropy—computed by applying a softmax over the embedding feature dimension—serves as a proxy for uncertainty in the embedding representation. To effectively model the hierarchical and sequential structure of speech, we adopt the scalar long short-term memory (sLSTM) network [25]. The sLSTM employs scalar gating mechanisms that optimizes the memory mixing while maintaining the ability to capture long-range dependencies in sequential data.

Building on this foundation, we further extend our work by proposing an improved multi-target intelligibility prediction model (iMTI-Net), which enhances the original MTI-Net framework by integrating uncertainty-aware features and

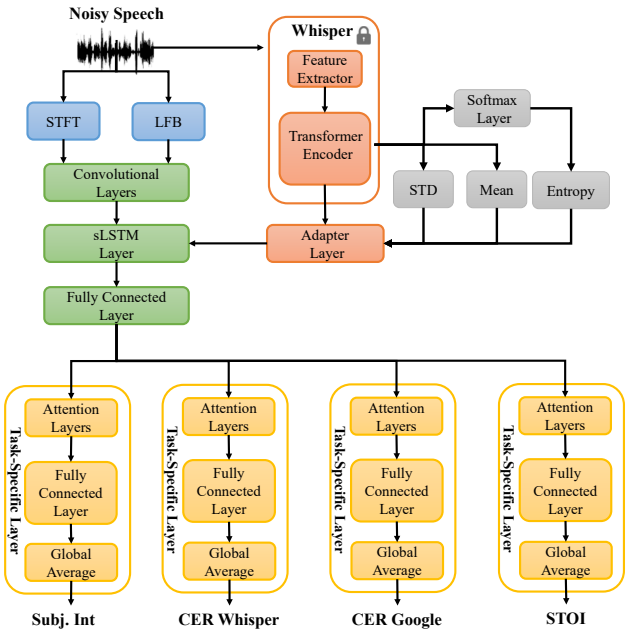


Fig. 1. Architecture of iMTI-Net.

a convolutional neural network (CNN) with sLSTM model. The proposed model employs a multi-task learning strategy to simultaneously predict both human and machine intelligibility scores. Machine intelligibility scores are represented by character error rate (CER) from two automatic speech recognition (ASR) systems, namely Google ASR [26] and Whisper [23], while human intelligibility scores include subjective listening test scores and objective metrics such as STOI. Experimental results demonstrate that iMTI-Net consistently outperforms the original MTI-Net across nearly all evaluation metrics, highlighting the effectiveness of incorporating uncertainty-aware features and the CNN-sLSTM architecture. Specifically, for intelligibility prediction, the iMTI-Net with CNN-sLSTM achieves the highest linear correlation coefficient (LCC) (0.7817) and Spearman's rank correlation coefficient (SRCC) (0.7622). For Whisper CER prediction, the CNN-BLSTM variant obtains the highest LCC (0.8151), while CNN-sLSTM achieves the best SRCC (0.8222) and MSE (0.0360). In the case of Google CER prediction, CNN-sLSTM outperforms all other models with the highest LCC (0.8505), SRCC (0.8403), and lowest MSE (0.0312). Similarly, for STOI prediction, CNN-sLSTM achieves the best LCC (0.9051) and SRCC (0.9150), while CNN-BLSTM obtains the lowest mean square error (MSE) (0.0031). These results confirm that the use of sLSTM, in combination with CNN and statistical features, leads to consistent and significant improvements in both human and machine intelligibility prediction tasks.

II. Proposed Method

The overall architecture of the iMTI-Net model is illustrated in Fig. 1 and comprises three feature extraction modules. First, the speech waveform $\boldsymbol{Y}$ undergoes short-time Fourier transform (STFT) to extract spectral features. Second, $\boldsymbol{Y}$ is processed through learnable filter banks (LFB) within a sinc-based convolutional network [27], producing complementary acoustic features. These two feature sets are concatenated along the feature dimension and fed into CNN layers; the output is denoted as $\boldsymbol{C}$.

The third module processes the same waveform $\boldsymbol{Y}$ using the Whisper model. Instead of using the Whisper embeddings directly as in prior work [19], our method extracts proxy uncertainty from these embeddings to provide richer information for intelligibility prediction. Specifically, we calculate the mean, standard deviation, and entropy of the embedding vectors at each time frame, which serve as proxies for the confidence and variability inherent in the embeddings. Formally, this is defined as:

$$
\begin{aligned}
\boldsymbol{E} &= \text{Whisper}(\boldsymbol{Y}) \in \mathbb{R}^{T\times D} \\
\mu_t &= \frac{1}{D}\sum_{d=1}^{D} E_{t,d} \\
\sigma_t &= \sqrt{\frac{1}{D}\sum_{d=1}^{D}(E_{t,d}-\mu_t)^2} \\
\boldsymbol{p}_t &= \text{softmax}(\boldsymbol{E}_t) \\
h_t &= -\sum_{d=1}^{D} p_{t,d}\log p_{t,d} \\
\boldsymbol{x}_t &= [\,\boldsymbol{E}_t; \mu_t; \sigma_t; h_t\,] \in \mathbb{R}^{D+3} \\
\tilde{\boldsymbol{x}}_t &= [\,\text{Adapter}(\boldsymbol{x}_t); \boldsymbol{C}_t\,]
\end{aligned}
\tag{1}
$$

where $\boldsymbol{E} \in \mathbb{R}^{T\times D}$ denotes the Whisper embeddings with T frames and D dimensions. μ_t, σ_t, and h_t represent the per-frame mean, standard deviation, and entropy, respectively, computed over the embedding dimensions. The feature vector $\boldsymbol{x_t}$ is formed by concatenating the original embedding $\boldsymbol{E}_t$ with its corresponding statistical features. Finally, $\tilde{\boldsymbol{x}}_t$ denotes the final feature representation, formed by concatenating the output of the adapter layer applied to $\boldsymbol{x}_t$ with the CNN-based acoustic feature $\boldsymbol{C}_t$.

Next, the feature representation $\tilde{\boldsymbol{x}}_t$ is processed through a sLSTM [25] network, which models temporal dependencies in the data. The sLSTM updates its internal state and output at each time step t based on the current input $\tilde{\boldsymbol{x}}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$. The computations are defined as follows:

$$
\begin{aligned}
\tilde{\boldsymbol{z}}_t &= \boldsymbol{w}_z^\top \tilde{\boldsymbol{x}}_t + \boldsymbol{r}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z, & \boldsymbol{z}_t &= \phi(\tilde{\boldsymbol{z}}_t) \\
\tilde{\boldsymbol{i}}_t &= \boldsymbol{w}_i^\top \tilde{\boldsymbol{x}}_t + \boldsymbol{r}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i, & \boldsymbol{i}_t &= \exp(\tilde{\boldsymbol{i}}_t) \\
\tilde{\boldsymbol{f}}_t &= \boldsymbol{w}_f^\top \tilde{\boldsymbol{x}}_t + \boldsymbol{r}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f, & \boldsymbol{f}_t &= \begin{cases} \exp(\tilde{\boldsymbol{f}}_t) \\ \sigma(\tilde{\boldsymbol{f}}_t) \end{cases} \\
\tilde{\boldsymbol{o}}_t &= \boldsymbol{w}_o^\top \tilde{\boldsymbol{x}}_t + \boldsymbol{r}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o, & \boldsymbol{o}_t &= \sigma(\tilde{\boldsymbol{o}}_t) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot \boldsymbol{z}_t \\
\boldsymbol{n}_t &= \boldsymbol{f}_t \cdot \boldsymbol{n}_{t-1} + \boldsymbol{i}_t \\
\tilde{\boldsymbol{h}}_t &= \frac{\boldsymbol{c}_t}{\boldsymbol{n}_t}, & \boldsymbol{h}_t &= \boldsymbol{o}_t \cdot \tilde{\boldsymbol{h}}_t
\end{aligned}
\tag{2}
$$

where, $\boldsymbol{w}_*$, $\boldsymbol{r}_*$, and $\boldsymbol{b}_*$ are input weights, recurrent weights, and biases corresponding to each gate and cell units. The functions $\phi(\cdot)$ and $\sigma(\cdot)$ denote the hyperbolic tangent and sigmoid activation functions, respectively.

Furthermore, compared to the standard LSTM, the sLSTM maintains an additional normalization state $\boldsymbol{n}_t$ alongside the cell state $\boldsymbol{c}_t$, allowing it to compute a normalized hidden state $\tilde{\boldsymbol{h}}_t = \frac{\boldsymbol{c}_t}{\boldsymbol{n}_t}$. This normalization helps stabilize training by preventing unbounded growth of the cell memory and improves the model's ability to capture long-range dependencies efficiently. The output of the sLSTM is further processed by a fully connected layer. Task-specific layers are then employed to predict each corresponding intelligibility metric. Finally, the iMTI-Net framework integrates both frame-level and utterance-level scores into the objective function for each metric's loss.

$$L = \gamma_1 L_{Int.} + \gamma_2 L_{CER_{ws}} + \gamma_3 L_{CER_{goo}} + \gamma_4 L_{STOI} \tag{3}$$

where weights between Intelligibility, CER Whisper, CER Google, and STOI are determined by γ_1, γ_2, γ_3, and γ_4, respectively.

III. Experiments

A. Experimental setup

The iMTI-Net model was evaluated using the TMHINT-QI(S) dataset [19], an extended version of the original TMHINT-QI corpus [28]. This extension incorporates additional unseen noise types, speakers, and speech enhancement models. Notably, TMHINT-QI(S) also serves as one of the benchmark tracks in the VoiceMOS Challenge 2023 [29]. The evaluation set consists of clean, noisy, and enhanced speech samples, including three seen noise types (babble, white, and pink) and one unseen noise condition (street noise). It also covers three seen enhancement systems—minimum mean square error (MMSE) [30], fully convolutional network (FCN) [31], and transformer [32]—alongside two unseen systems: conformer-based metric generative adversarial network (CM-GAN) [33] and DEMUCS [34]. In total, the evaluation set consists of 1,960 utterances with a quality score (ranging from 0 to 5) and an intelligibility score (ranging from 0 to 1). Additionally, we prepared CER scores from two ASR systems—Google ASR [26] and Whisper [23]. For consistency across metrics, the CER scores are inverted so that higher scores reflect better recognition performance.

To evaluate prediction performance, we adopt three commonly used metrics: LCC, SRCC [35], and MSE. Both LCC and SRCC assess the strength of the relationship between the predicted and ground-truth scores, with higher values indicating better alignment and overall performance. On the other hand, lower MSE values indicate better performance.

B. Performance comparison between baseline and iMTI-Net

In this experiment, we compare the performance of iMTI-Net with the original MTI-Net [15]. While we follow the original implementation of MTI-Net for model deployment, we introduce slight modifications by incorporating additional assessment metrics during training. The original MTI-Net was trained to predict Intelligibility, Google CER, and STOI. In our setup, we also include Whisper CER as an additional target. Furthermore, given the strong performance of the Whisper model, we replace HuBERT with Whisper in our implementation. For simplicity, we refer to this modified version of MTI-Net as the baseline.

TABLE I
LCC, SRCC, AND MSE RESULTS BETWEEN BASELINE AND IMTI-NET FOR SPEECH INTELLIGIBILITY PREDICTION.

	Architecture	LCC	SRCC	MSE
Baseline	CNN-BLSTM	0.7630	0.7071	0.0249
iMTI-Net	CNN-BLSTM	0.7791	0.7581	0.0262
iMTI-Net	CNN-sLSTM	**0.7817**	**0.7622**	**0.0259**

TABLE II
LCC, SRCC, AND MSE RESULTS BETWEEN BASELINE AND IMTI-NET FOR CER OF WHISPER PREDICTION.

	Architecture	LCC	SRCC	MSE
Baseline	CNN-BLSTM	0.7118	0.6600	0.0471
iMTI-Net	CNN-BLSTM	**0.8151**	0.8145	0.0334
iMTI-Net	CNN-sLSTM	0.8031	**0.8222**	**0.0360**

For the iMTI-Net model, we deploy two variants: one using a CNN-BLSTM architecture and the other using a CNN-sLSTM architecture. The CNN-BLSTM model consists of 12 convolutional layers with four channel groups (16, 32, 64, and 128 channels), followed by a one-layer BLSTM with 128 units and a fully connected layer with 128 neurons. Four prediction branches are used, each comprising an attention layer, a fully connected layer with a single output neuron, and a global average pooling operation to generate the predicted quality and intelligibility scores. During training, we set the loss weights as $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 1$, and $\gamma_4 = 5$, and use a learning rate of 1e-5. Unlike the baseline model, which concatenates features along the temporal dimension, our iMTI-Net performs feature concatenation along the feature dimension. Additionally, the corresponding statistical features are concatenated with the original features in all iMTI-Net model variants. Lastly, the CNN-sLSTM version follows the same configuration, except that the sLSTM replaces the BLSTM component, while retaining the same architectural structure.

Experimental results confirm that the proposed iMTI-Net consistently outperforms the Baseline across all evaluation metrics. As shown in Table 1, both iMTI-Net variants yield better performance for intelligibility prediction, with the CNN-sLSTM model achieving the highest LCC (0.7817) and SRCC (0.7622), along with a competitive MSE (0.0259). This highlights the effectiveness of combining statistical features with the sLSTM architecture in modeling intelligibility-related patterns. For Whisper CER prediction (Table 2), the iMTI-Net with CNN-BLSTM achieves the highest LCC (0.8151), while the CNN-sLSTM variant delivers the best SRCC (0.8222) and MSE (0.0360). In Table 3, the iMTI-

TABLE III
LCC, SRCC, AND MSE RESULTS BETWEEN BASELINE AND IMTI-NET FOR CER OF GOOGLE PREDICTION.

	Architecture	LCC	SRCC	MSE
Baseline	CNN-BLSTM	0.8156	0.7391	0.0373
iMTI-Net	CNN-BLSTM	0.8443	0.8358	0.0323
iMTI-Net	CNN-sLSTM	**0.8505**	**0.8403**	**0.0312**

TABLE IV
LCC, SRCC, AND MSE RESULTS BETWEEN BASELINE AND IMTI-NET FOR STOI PREDICTION.

	Architecture	LCC	SRCC	MSE
Baseline	CNN-BLSTM	0.8693	0.8893	0.0060
iMTI-Net	CNN-BLSTM	0.9005	0.9050	**0.0031**
iMTI-Net	CNN-sLSTM	**0.9051**	**0.9150**	0.0039

Net with CNN-sLSTM again leads across all metrics for Google CER prediction—LCC (0.8505), SRCC (0.8403), and MSE (0.0312)—indicating that sLSTM offers a more effective sequential modeling framework than BLSTM in this context. Table 4 shows the results for STOI prediction, where iMTI-Net with CNN-sLSTM achieves the highest LCC (0.9051) and SRCC (0.9150), while iMTI-Net with CNN-BLSTM obtains the lowest MSE (0.0031). These findings further demonstrate the robustness and effectiveness of the proposed iMTI-Net, particularly with the CNN-sLSTM configuration.

C. Qualitative analysis between baseline and iMTI-Net

In this experiment, we focus on a qualitative comparison between the Baseline and iMTI-Net models. For iMTI-Net, we select the best-performing variant that uses CNN-sLSTM. The scatter plots, as shown in Fig.2, show that the Baseline tends to cluster predictions in the mid-range and struggles to predict very low or high intelligibility scores. In contrast, iMTI-Net produces a more even spread of predictions across the full range, showing better sensitivity to both low and high intelligibility. This suggests that the CNN-sLSTM structure and uncertainty-awareness help the model better capture variations in speech, leading to more accurate predictions.

IV. CONCLUSIONS

This work presents iMTI-Net, an improved multi-target model for non-intrusive speech intelligibility prediction. iMTI-Net integrates uncertainty-aware statistical features derived from Whisper embeddings with a CNN-sLSTM architecture and is trained using a multi-task learning strategy to jointly predict human intelligibility scores and machine-based CER from Google ASR and Whisper. To capture global speech characteristics and proxy uncertainty, we incorporate the mean, standard deviation, and entropy of the Whisper embeddings. Experimental results confirm that iMTI-Net consistently outperforms the original MTI-Net across intelligibility, STOI, and CER prediction tasks. The iMTI-Net with CNN-sLSTM variant achieves the best performance in most metrics, confirming the benefit of sLSTM in optimizing memory mixing while maintaining the ability to capture long-range dependencies in sequential data. In future work, we plan to explore broader use of uncertainty-aware representations and extend iMTI-Net to handle more diverse and unseen speech scenarios.

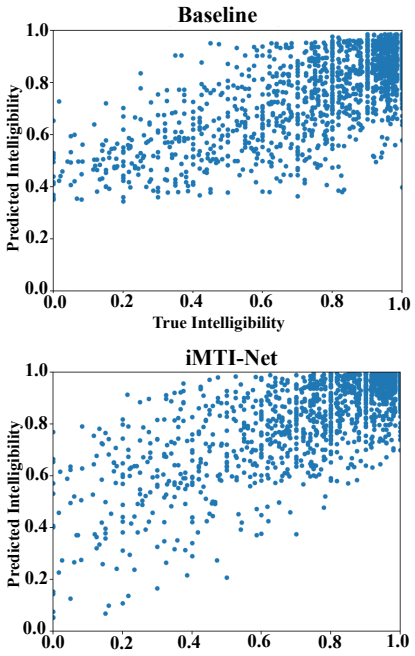


Fig. 2. Scatter plots of Baseline and iMTI-Net for predicting subjective intelligibility.

REFERENCES

[1] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC press, 2007.

[2] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," in *ITU-T Recommendation*, 2001, p. 862.

[3] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[4] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[5] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[6] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[7] ANSI Std. S3.5 1997, "Methods for calculation of the speech intelligibility index," in *Acoustical Society of America*, 1997.

[8] T. Houtgast and H. 1. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *Proc. ICASSP*, 2020, pp. 336–340.

[11] A. H. Andersen, J. M. D. Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

[12] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A deep learning-based non-intrusive speech intelligibility assessment model," in *Proc. APSIPA ASC*, 2020, pp. 482–486.

[13] H.-T. Chiang, Y.-C. Wu, C. Yu, *et al.*, "HASA-Net: A non-intrusive hearing-aid speech assessment network," in *Proc. ASRU*, 2021, pp. 907–913.

[14] Z. Tu, N. Ma, and J. Barker, "Exploiting hidden representations from a DNN-based speech recogniser for speech intelligibility prediction in hearing-impaired listeners," in *Proc. INTERSPEECH*, 2022, pp. 3488–3492.

[15] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MTI-Net: A multi-target speech intelligibility prediction model," in *Proc. INTERSPEECH*, 2022, pp. 5463–5467.

[16] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids," in *Proc. INTERSPEECH*, 2022, pp. 3944–3948.

[17] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *Proc. ICASSP*, 2024, pp. 1421–1425.

[18] R. Mogridge, G. Close, R. Sutherland, *et al.*, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," in *Proc. ICASSP*, 2024, pp. 306–310.

[19] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," in *Proc. ICME*, 2024, pp. 1–6.

[20] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss," *Applied Acoustics*, vol. 214, p. 109 663, 2023.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, 2020, pp. 1–12.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 3451–3460, 2021.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[24] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wan, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[25] M. Beck, K. Pöppel, M. Spanring, *et al.*, "XLSTM: Extended long short-term memory," in *Proc. NeurIPS*, 2024, pp. 1–57.

[26] A. Zhang, "Speech recognition (version 3.6) [software]," in *Proc. ICCC*, 2017.

[27] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with Sincnet," in *Proc. SLT*, 2018, pp. 1021–1028.

[28] Y.-W. Chen and Y. Tsao, "InQSS: A speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. INTERSPEECH*, 2022, pp. 3088–3092.

[29] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *Proc. ASRU*, 2023, pp. 1–7.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[31] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA ASC*, 2017, pp. 6–12.

[32] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6649–6653.

[33] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based MetricGAN for speech enhancement," in *Proc. INTERSPEECH*, 2022, pp. 936–940.

[34] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv 1911.13254*, 2021.

[35] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.